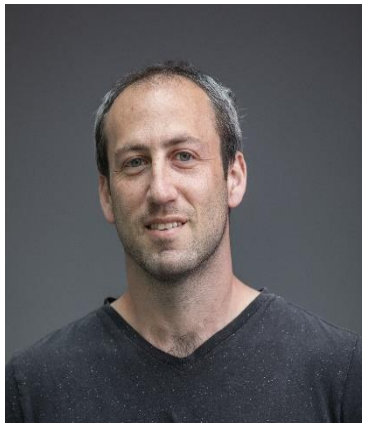
Is this correct? Let's Check!

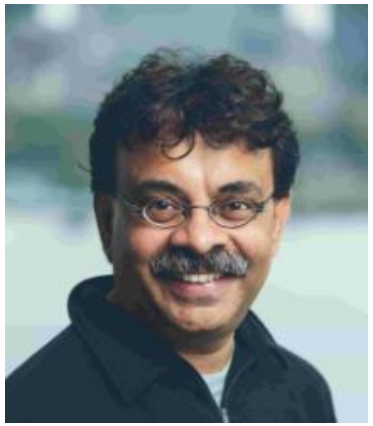
Omri Ben-Eliezer

Dan Mikulincer



Elchanan Mossel

Madhu Sudan



Is societal knowledge robust?

- Why ask this question?
  - Builds on error-prone processes
    - Collecting Data
    - Analyzing it
    - Combining results
- Last is especially problematic/interesting: Knowledge is cumulative!!
  - Accumulation can be very bad for errors!!!!!
- There must exist error-correcting processes
  - What are they? How do they work? How well do they work?

Lebesgue's Mistake

- In 1904 Lebesgue proved the following theorem:

  "A projection of a measurable set is measurable"

Sur les fonctions représentables analytiquement ;

PAR M. H. LEBESGUE.

I. — Introduction.

Bien que, depuis Dirichlet et Riemann, on s'accorde généralement à dire qu'il y a fonction quand il y a correspondance entre un nombre y et des nombres $x_1, x_2, \ldots, x_n$, sans se préoccuper du procédé qui sert à établir cette correspondance, beaucoup de mathématiciens semblent ne considérer comme de vraies fonctions que celles qui sont établies par des correspondances analytiques. On peut penser qu'on introduit peut-être ainsi une restriction assez arbitraire ; cependant il est certain que cela ne restreint pas pratiquement le champ des applications, parce que, seules, les fonctions représentables analytiquement sont effectivement employées jusqu'à présent.

Lebesgue's Mistake

- In 1904 Lebesgue proved the following theorem:

  "A projection of a measurable set is measurable"

- According to Google Scholar the paper has 303 citations.
- Some citations prior to 1917.

[CITATION] **Sur** les **fonctions représentables analytiquement**
H Lebesgue - Journal de mathematiques pures et appliquees, 1905 - eudml.org
EUDML | **Sur** les **fonctions représentables analytiquement** ... **Sur** les **fonctions représentables analytiquement** ... **Sur** les **fonctions représentables analytiquement**." ...
☆ Save 99 Cite Cited by 303 Related articles All 2 versions »

Sulla rappresentazione analitica delle funzioni di piÙ variabili reali

Leonida Tonelli

Rendiconti del Circolo Matematico di Palermo (1884-1940) **29**, 1–36 (1910) | Cite this article

References

2) *Sur Us fonctions représentables analytiquement* [Journal de Mathématiques pures et appliquées, VIe série, I (1905), pp. 139–216].

Lebesgue's Mistake

- In 1904 Lebesgue proved the following theorem:

  "A projection of a measurable set is measurable"

- According to Google Scholar the paper has 303 citations.
- Some citations prior to 1917.
- In 1917 Suslin discovered a counterexample:

  "There exists a projection of measurable set which is not mesurable"

- Happy ending: The field of "Descriptive set theory" was born.
- **Did the mistake propagate?**

Another Example

The NEW ENGLAND JOURNAL of MEDICINE

Dietary Fats, Carbohydrates and Atherosclerotic Vascular Disease

•List of authors.Robert B. McGandy, M.D.‡,

•D.M. Hegsted, Ph.D.‡,

•and F. J. Stare, M.D.§

1967

Dietary fats, carbohydrates and atherosclerotic vascular disease
RB McGandy, DM Hegsted… - New England Journal of …, 1967 - Mass Medical Soc
THERE is considerable evidence relating nutrition, presumably through its influence on the
levels of circulating lipids, to the relentless progression of atherosclerotic vascular disease …
Save Cite Cited by 180 Related articles

E.G:

Epidemiology as a guide to clinical decisions
SB Hulley, RH Rosenman, RD Bawol, RJ Brand - Nutrition Today, 1980 - journals.lww.com
Epidemiology Page 1 Epidemiology aS a Guide tO Clinical Decisions The association between Triglyceride and coronary heart disease is discussed in this report of work supported by a …
Save Cite Cited by 965 Related articles All 13 versions

In the 1960s, the sugar industry funded research that downplayed the risks of sugar and highlighted the hazards of fat,
according to a newly published article in *JAMA Internal Medicine*.

Special Communication
November 2016
Sugar Industry and Coronary Heart Disease ResearchA Historical Analysis of Internal Industry Documents
Cristin E. Kearns, DDS, MBA[1,2]; Laura A. Schmidt, PhD, MSW, MPH[1,3,4]; Stanton A. Glantz, PhD[1,5,6,7,8]

Today

Question

Can we guarantee that the effects caused by a single error do not propagate?

Sur les fonctions représentables analytiquement ;

PAR M. H. LEBESGUE.

I. — Introduction.

Bien que, depuis Dirichlet et Riemann, on s'accorde généralement à dire qu'il y a fonction quand il y a correspondance entre un nombre y et des nombres $x_1, x_2, \ldots, x_n$, sans se préoccuper du procédé qui sert à établir cette correspondance, beaucoup de mathématiciens semblent ne considérer comme de vraies fonctions que celles qui sont établies par des correspondances analytiques. On peut penser qu'on introduit peut-être ainsi une restriction assez arbitraire ; cependant il est certain que cela ne restreint pas pratiquement le champ des applications, parce que, seules, les fonctions représentables analytiquement sont effectivement employées jusqu'à présent.

[CITATION] **Sur les fonctions représentables analytiquement**

H Lebesgue - Journal de mathematiques pures et appliquees, 1905 - eudml.org

EUDML | **Sur les fonctions représentables analytiquement** ... **Sur les fonctions représentables analytiquement** ... **Sur les fonctions représentables analytiquement**." ...

☆ Save 99 Cite Cited by 303 Related articles All 2 versions »

Sulla rappresentazione analitica delle funzioni di piÙ variabili reali

Leonida Tonelli

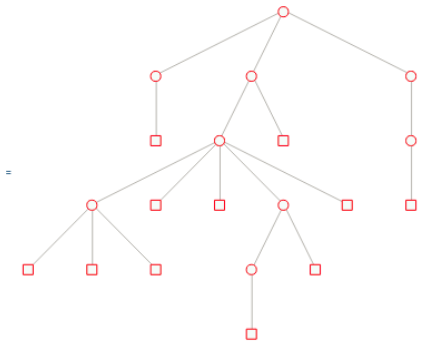
Rendiconti del Circolo Matematico di Palermo **29**, 1–36 (1910) Cite this article

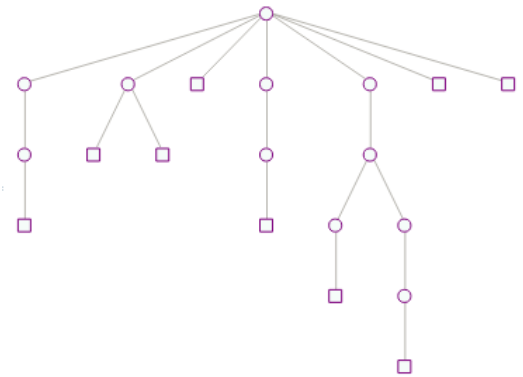
References

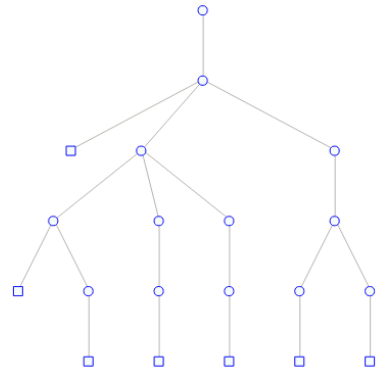
2) *Sur Us fonctions représentables analytiquement* [Journal de Mathématiques pures et appliquées, VIe série, I (1905), pp. 139–216].

Cumulative Knowledge Process

- In the paper we model the process of accumulating knowledge.
- Main properties:
  - New “units of knowledge” build upon previous units.
  - Errors are sometimes introduced and may propagate forward.
  - Errors can be checked and removed from the process.
- We study **structural properties** of the process.



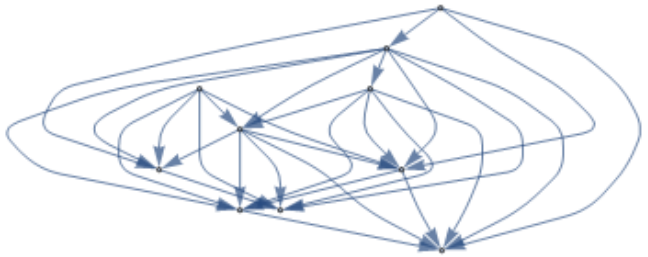




The Model

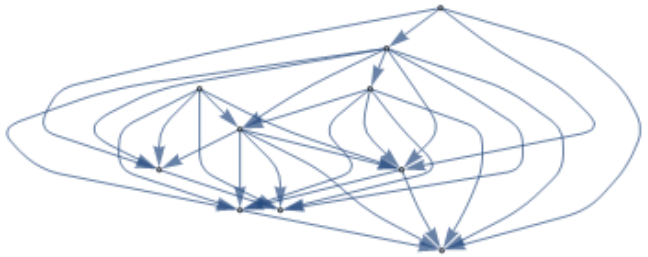
Representation of Knowledge

- **Ideally**: Knowledge is stored as Directed Acyclic Graph (DAG).
  - Vertices represent units of knowledge
  - Edges represent dependence or “inherited knowledge”.
  - E.g. a paper cites several papers.



Representation of Knowledge

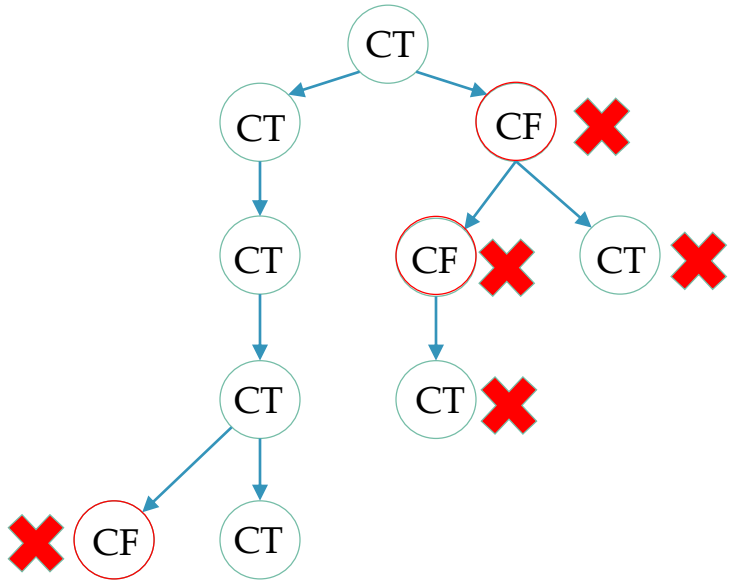
- **Ideally**: Knowledge is stored as Directed Acyclic Graph (DAG).
  - Vertices represent units of knowledge
  - Edges represent dependence or "inherited knowledge".
  - E.g. a paper cites several papers.



- Citation correlation? Would require a proper model of knowledge clustering.
- **Simplified notion:** Knowledge is represented as a tree.

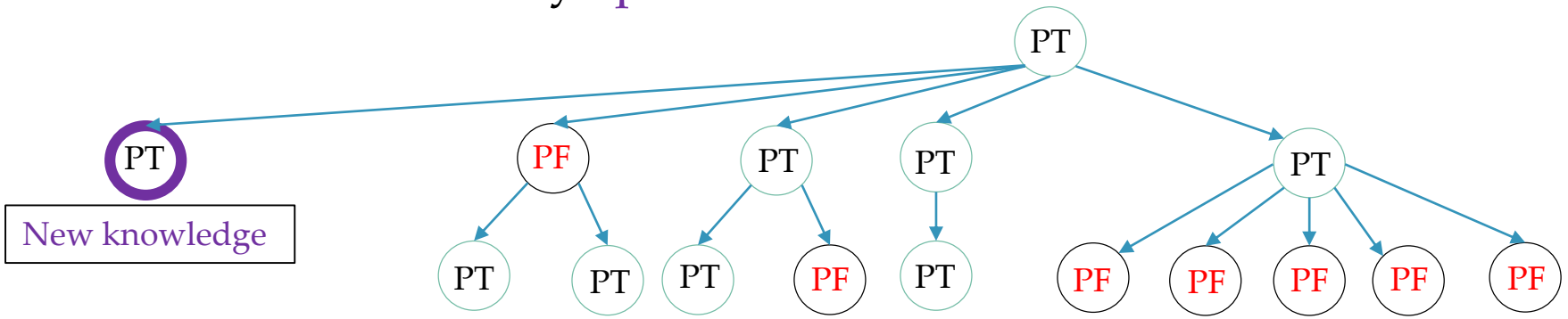
The Model

- **The Knowledge ~~DAG~~ Tree**: In the Cumulative Knowledge Process (CKP) knowledge units are modeled as a tree.
- A **node** represents a single "unit of knowledge" and **edges** represent the relation of "building upon existing knowledge"
- Each node has:
  - A hidden state conditionally true(CT)/conditionally false(CF)
  - A public state proclaimed true(PT)/proclaimed false(PF)
  - A node is considered to hold true knowledge
    if **all ancestors are CT**



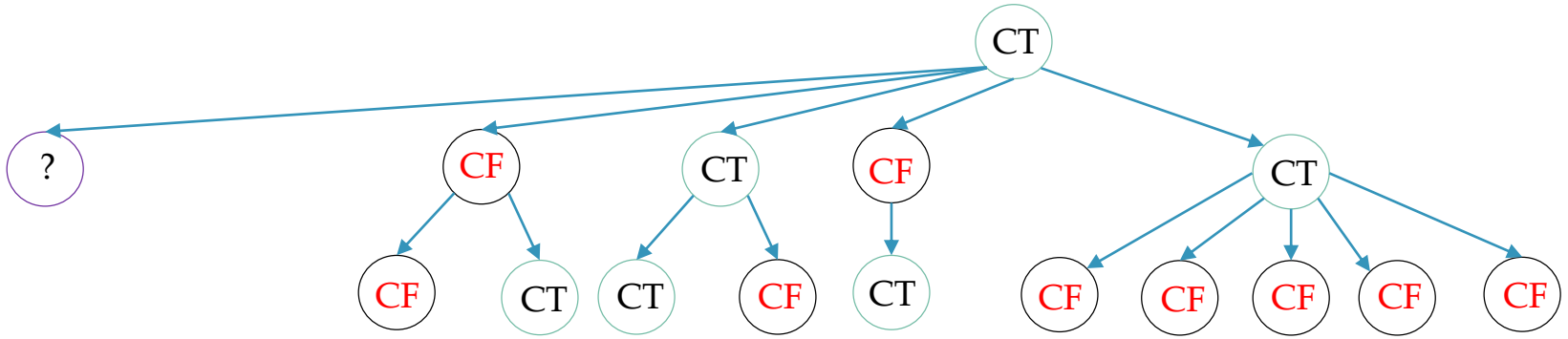
Accumulating Knowledge

- At each time $t \geq 0$, we have a knowledge tree T_t with associated labels.
- At time $t + 1$ a new node is added to the tree, by choosing a **random** proclaimed true (PT) parent.
- Parents are chosen according to the preferential attachment model.
  - **The more PT children a node has the more likely it is to generate new knowledge.**
- A new node is always proclaimed true



Injection of Errors

- Recall: nodes also have hidden states.
- The hidden label of a new node is determined randomly:
  - Parameter ε: with probability ε the new node is CF and otherwise CT.

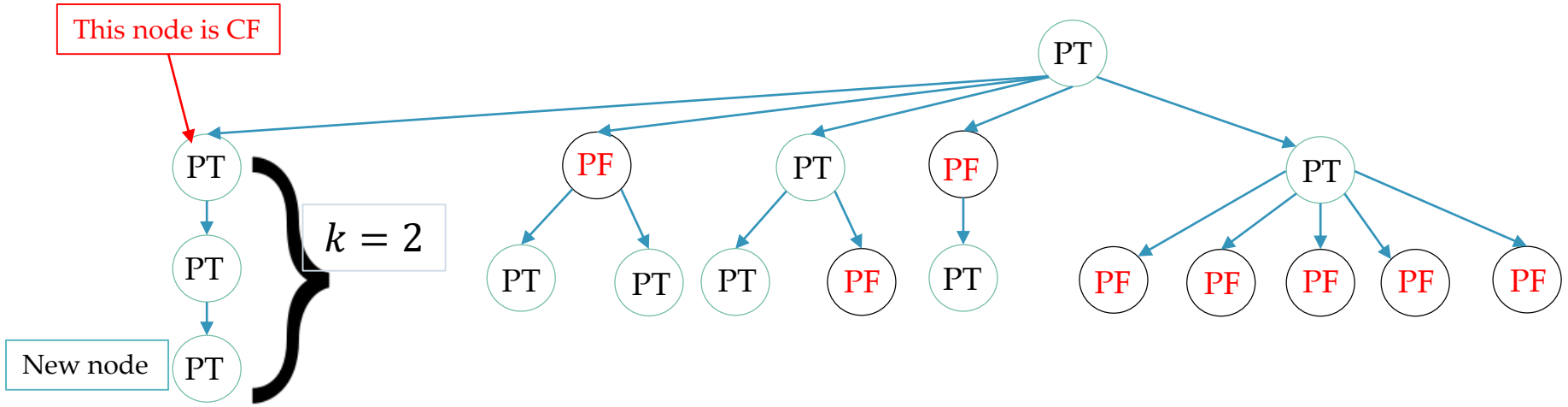


Aside: Probabilistic Models

- Quote from unknown source*
  - "All models are wrong. Some are useful"
- Obviously today's model is too simple to model the complex phenomena
  - Question (for you) to ponder : Is it useful?

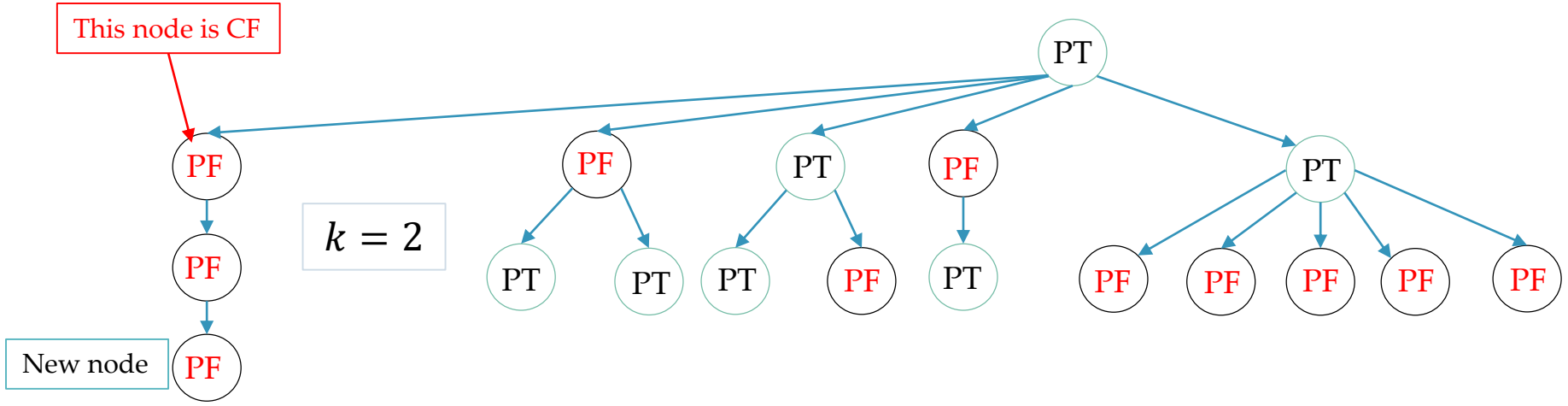
Checking for Errors

- Checks may be performed whenever a new node is added.
  - Parameter p: a node preformed a check with probability p.
- Checks are performed by ascending the tree.
  - Parameter k: the number of levels to be checked.



Checking for Errors

- Checks may be performed whenever a new node is added.
  - Parameter p: a node preformed a check with probability p.
- Checks are performed by ascending the tree.
  - Parameter k: the number of levels to be checked.
- If a CF or PF node is encountered, the public state of the **entire path** changes to proclaimed false.



The Model - Summary

- **Knowledge:** represented by a tree.
- **Growth:** by preferential attachment; nodes of high degree are more influential.
- **Errors:** introduced (sometimes) when new knowledge is created
- **Checks:** Performed when new knowledge is introduced, with some probability in locality of new knowledge.
- **Error correction:** when a node is verified to be faulty, the error is announced and the node is effectively eliminated.

With the parameters $\varepsilon, p,$ and k the model is called the $(\boldsymbol{\varepsilon}, \boldsymbol{p}, \boldsymbol{k}) - \boldsymbol{CKP}$.

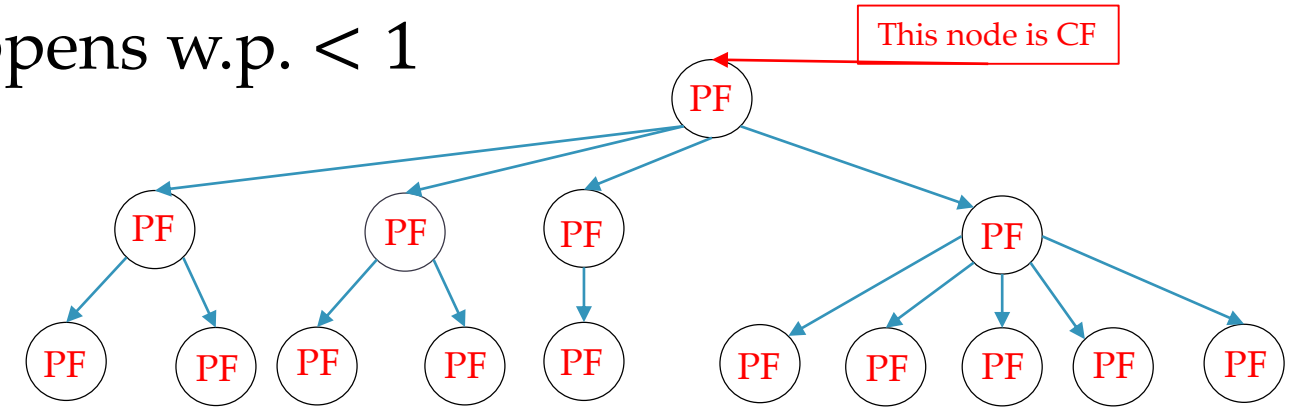
Phenomena

Error Effects

Question

Can we guarantee that the effects caused by a single error do not propagate?

- **Effects caused by a single error:** subtree rooted at a CF node.
- **Observation:** if, at some time, all nodes in a subtree are marked PF, the subtree is effectively eliminated for all future.
- **Error effect elimination:** If this happens w.p. 1
- **Error effect survives:** If this happens w.p. < 1



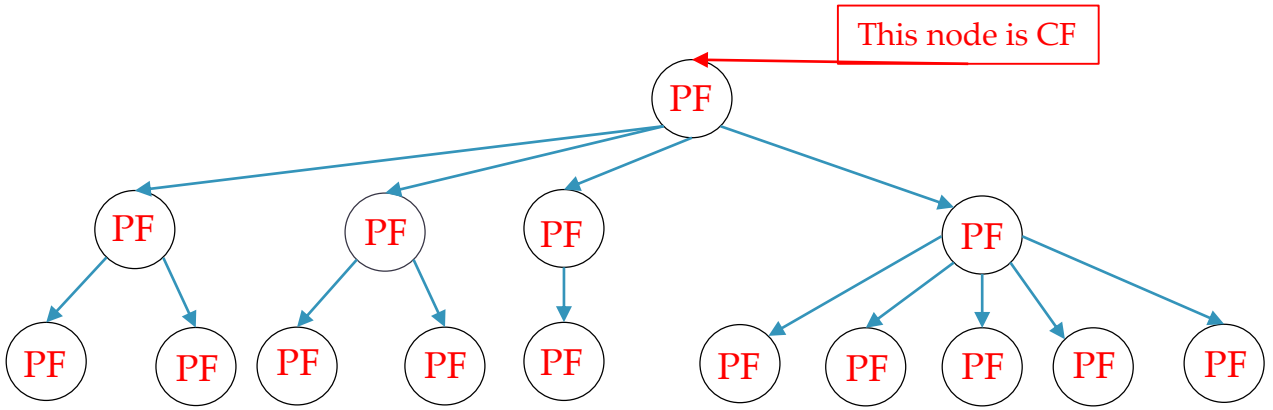
Error Effects

Question

Can we guarantee that the effects caused by a single error do not propagate?

Definition

- If every subtree rooted at a CF node is eliminated with probability 1, we say that the **error effects are completely eliminated.**
- Otherwise, the **error effects survive with positive probability.**



Phenomena of interest

- Do error effects survive? Or get eliminated?
- When the *error effect survives*:
  - How large are false components (compared to true ones)?
- When the *error effect is completely eliminated*:
  - How large are the "temporary error effects"?

Results

First Result – Depth Matters

Theorem 1

If $\boldsymbol{k = 2}$, then for any $\boldsymbol{p < 1}$, the error effects in the $\boldsymbol{(\varepsilon, p, 2) - CKP}$ survives with positive probability.

Second Result– Checking Matters

Theorem 2

For every $\boldsymbol{k \geq 4}$, and $\boldsymbol{\varepsilon} < 1$, there exists $p_0 \in \left[\frac{1}{4}, 1\right)$ such that:

- If $p > p_0$ the error effects in the $(\boldsymbol{\varepsilon}, \boldsymbol{p}, \boldsymbol{k}) - \boldsymbol{CKP}$ are completely eliminated.
- If $p < p_0$ the error effects in the $(\boldsymbol{\varepsilon}, \boldsymbol{p}, \boldsymbol{k}) - \boldsymbol{CKP}$ survive with positive probability.

Further Results

- With a refined analysis we also consider other structural properties.
- When the *error effect survives*:
  - Identify parameters which ensure that **false components are sublinear**.
  - Also **control the size** of the components.
- When the *error effect is completely eliminated*:
  - Identify parameters which also ensure that **proportion of false nodes** in the tree is always **at the noise level**.

Proofs?

- On board

Future Directions

- **The mysterious case of $k = 3$**:
  - Can the error effect be eliminated when only performing depth 3 checks?
- **Phase transitions**:
  - Determine the value of the critical probability p_0.
- **More general models**:
  - Can similar results be obtained for DAGS, instead of tree?
  - Will require to define an appropriate preferential attachment model on DAGs, which allows "similar knowledge" units to cluster.

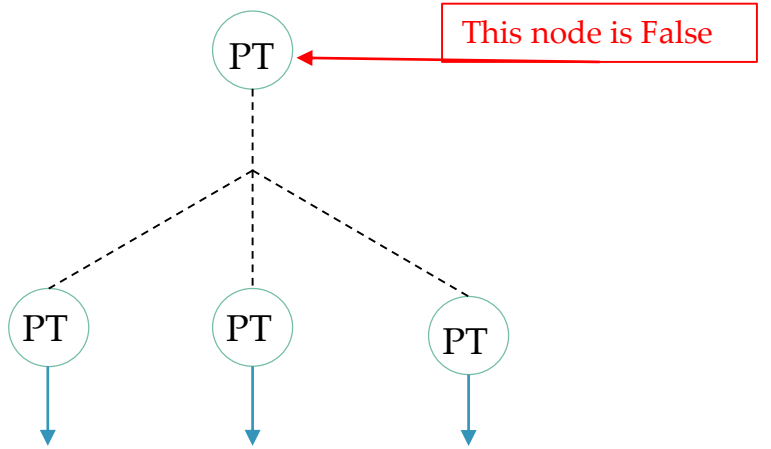
Thank you!

First Result – Depth Matters

Theorem 1

If $\boldsymbol{k = 2}$, then for any $\boldsymbol{p < 1}$, the error effects in the $\boldsymbol{(\varepsilon, p, 2) - CKP}$ survives with positive probability.

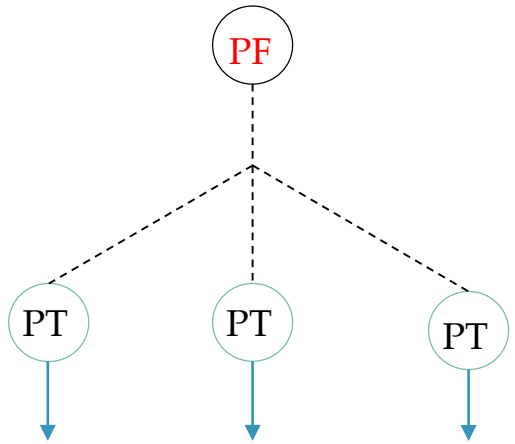
- Main idea: couple the $\boldsymbol{CKP}$ with a branching process.



First Result – Depth Matters

Theorem 1
If $\boldsymbol{k = 2}$, then for any $\boldsymbol{p < 1}$, the error effects in the $\boldsymbol{(\varepsilon, p, 2) - CKP}$ survive with positive probability.

- Main idea: couple the ***CKP*** with a branching process.
- We show that when $k = 2$, by the time an erroneous node is proclaimed false it will effectively create many new components.



First Result – Depth Matters

Theorem 1

If $\boldsymbol{k = 2}$, then for any $\boldsymbol{p < 1}$, the error effects in the $\boldsymbol{(\varepsilon, p, 2) - CKP}$ survive with positive probability.

Conclusion:

To guarantee that error effects are completely eliminated shallow checks are not enough!

Second Result– Main Ideas

- Proof of Theorem 2 is based on a (sub\super-)martingale analysis.
- We consider some observables in the process and identify regimes in which they increase or decrease in expectation.
- Examples:
  - Number of proclaimed true leaves in the tree.
  - Distribution of depths in proclaimed true subtree.